BỘ DỮ LIỆU GÁN NHÃN

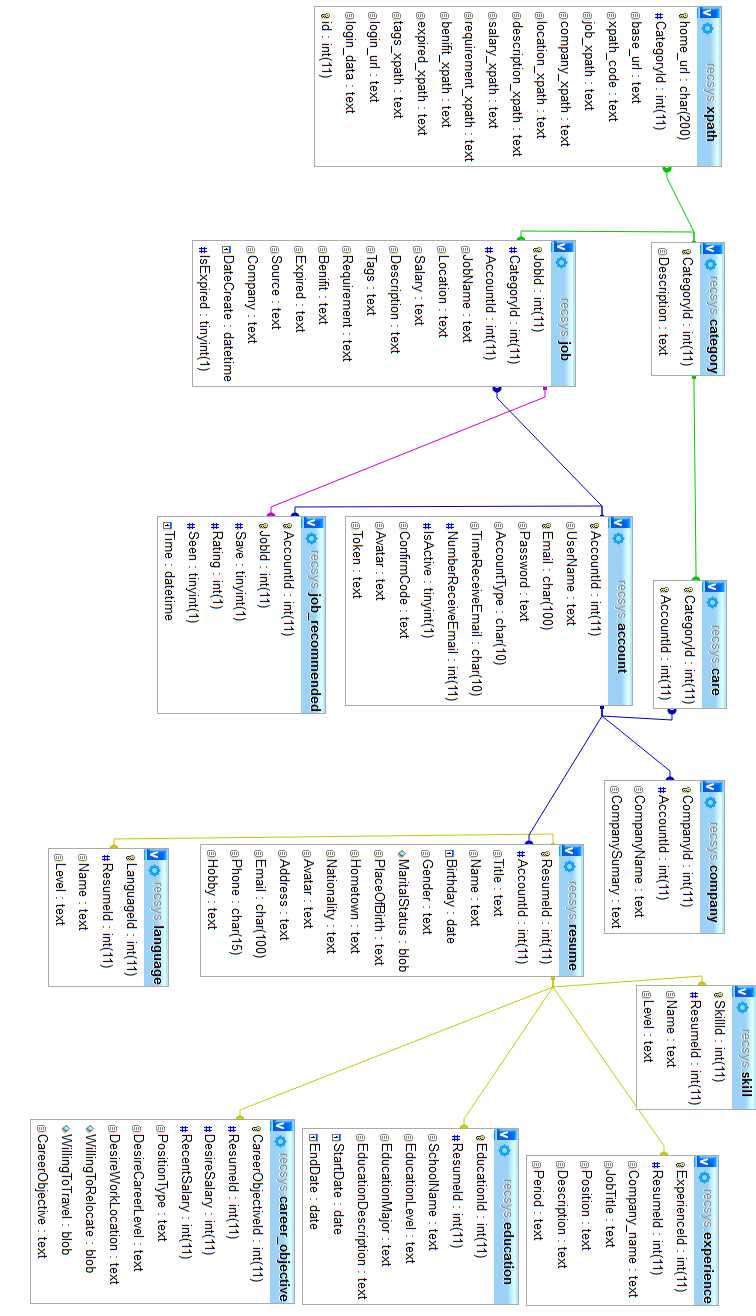
**Job Recommendation System**

# Thu thập dữ liệu

* Dữ liệu việc làm được thu thập từ 3 trang web việc làm lớn hiện nay ở Việt Nam gồm: vietnamworks.com, careerlink.vn và itviec.com.
* Việc thực hiện gán nhãn dữ liệu được thực hiện bởi nhóm nghiên cứu và một số bạn sinh viên hỗ trợ từ phía anh Dũng

# Tổ chức lưu trữ

Toàn bộ dữ liệu việc làm và dữ liệu đánh giá của người dùng, các bộ CV (résume) được tổ chức lưu trữ chung trong 1 database có cấu trúc như sau:

******

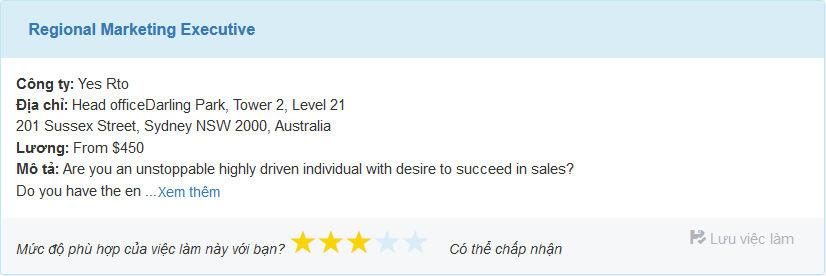
Sơ đồ 1. Lược đồ quan hệ

Một số bảng dữ liệu quan trọng cần lưu ý:

* Bảng job: nơi lưu trữ dữ liệu việc làm thu thập được từ công cụ thu thập dữ liệu
* Bảng account: lưu trữ thông tin tài khoản các người dùng của website
* Bảng job\_recommended: lưu trữ những việc làm đã được người dùng đánh giá
* Bảng resume: lưu trữ thông tin hồ sơ ứng viên (résume)

# Dữ liệu gán nhãn

* Bộ dữ liệu gán nhãn = việc làm + đánh giá của người dùng + CV của người dùng.
* Từ dữ liệu việc làm thu thập được, hệ thống gửi việc làm cho người dùng của hệ thống để người dùng đánh giá những việc làm đó. Việc làm được chọn để gửi gồm: 50% việc làm lĩnh vực công nghệ thông tin và 50% việc làm ngẫu nhiên từ các ngành nghề khác.
* Người dùng thực hiện đánh giá việc làm theo mức độ phù hợp của việc làm đó với nhu cầu của họ. Mức độ phù hợp của việc làm được đánh giá theo 5 cấp độ từ 1 - 5:
* 1 (Không liên quan): việc làm này hoàn toàn không liên quan đến nhu cầu của người đánh giá. Mức độ phù hợp từ 0 – 20%
* 2 (Không phù hợp): việc làm này có thể liên quan tới lĩnh vực ngành nghề của người đánh giá, nhưng nó không phù hợp với nhu cầu của họ. Mức độ phù hợp từ 20 – 40%
* 3 (Có thể chấp nhận): việc làm này tương đối phù hợp với nhu cầu của người đánh giá, mặc dù có thể nó không đáp ứng được một vài tiêu chí. Mức độ phù hợp từ 40 – 60%
* 4 (Khá phù hợp): việc làm này có mức độ phù hợp với người đánh giá khá cao, người đánh giá có khả năng cao sẽ apply vào vị trí việc làm này. Mức độ phù hợp từ 60 – 80%
* 5 (Rất phù hợp): việc làm này phù hợp hoàn toàn với người đánh giá về mọi tiêu chí. Mức độ phù hợp từ 80 – 100%

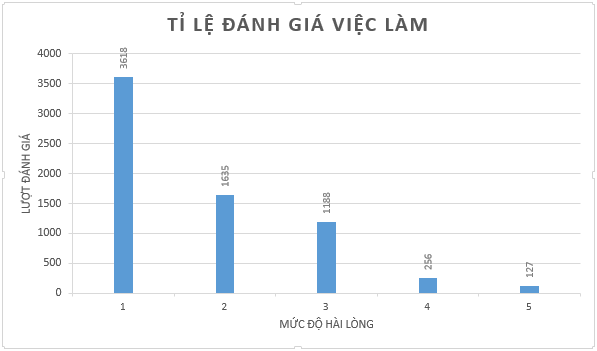


Hình 1. Cách đánh giá việc làm

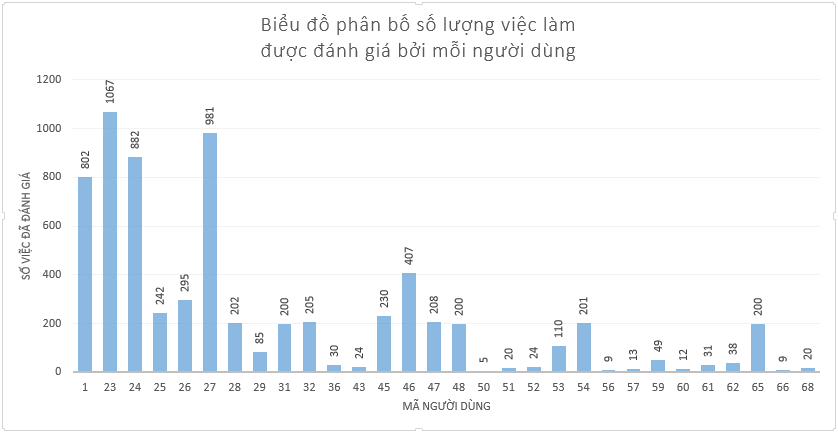
- CV của người dùng do chính người dùng website cung cấp, mỗi CV ứng với một mong muốn việc làm cụ thể ở một vài tiêu chí nào đó.

# Thống kê kết quả

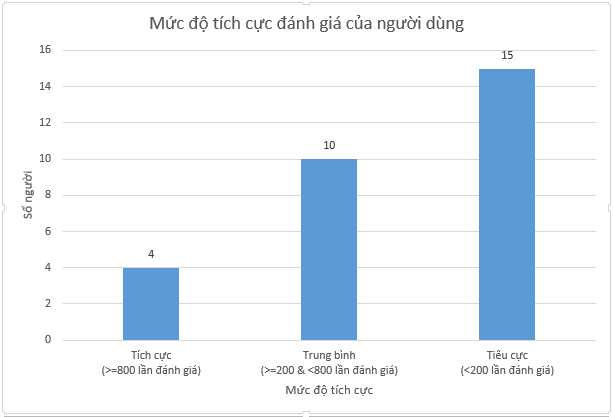
* Website thu thập dữ liệu việc làm:
  + <http://vietnamworks.com>
  + <http://careerlink.vn>
  + <http://itviec.com>
* Thời gian thu thập: từ 03/10/2015 đến 14/02/2016
* Số người dùng của website: 50
* Số CV (résume): 43
* Số việc làm: 39390.
* Số dữ liệu gán nhãn: 6824



Biểu đồ 1. Biểu đồ tỉ lệ đánh giá việc làm



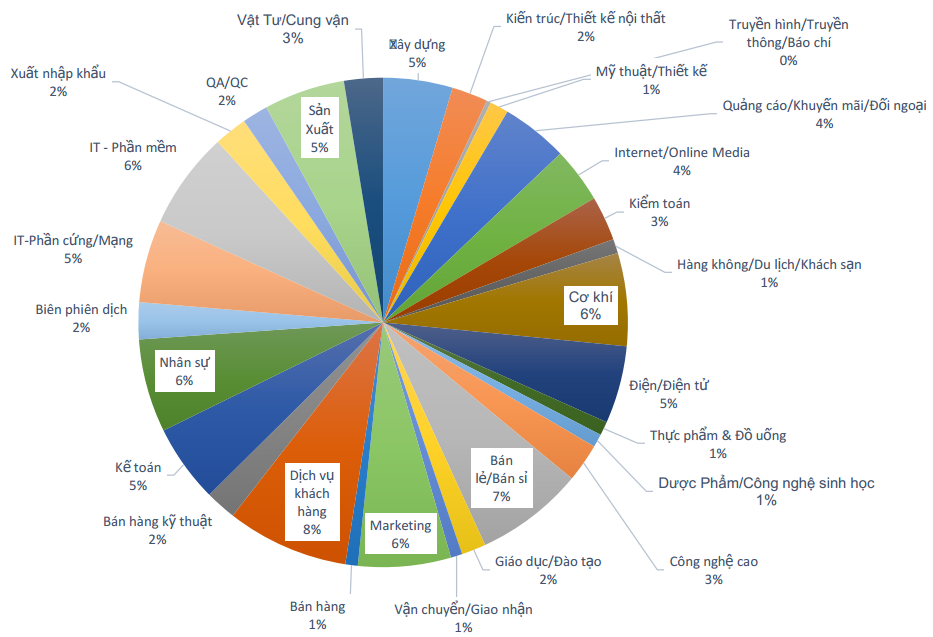
Biểu đồ 2. Biểu đồ phân bố số lượng việc làm được đánh giá bởi mỗi người dùng



Biểu đồ 3. Biểu đồ mức độ tích cực đánh giá của người dùng

* Việc làm thu thập được: 39390/ 31 ngành nghề

|  |  |  |
| --- | --- | --- |
| **Ngành nghề** | **Số việc làm** | **Tỉ lệ %** |
| ﻿Xây dựng | 1805 | 4.58 |
| Kiến trúc/Thiết kế nội thất | 937 | 2.38 |
| Truyền hình/Truyền thông/Báo chí | 112 | 0.28 |
| Mỹ thuật/Thiết kế | 487 | 1.24 |
| Quảng cáo/Khuyến mãi/Đối ngoại | 1733 | 4.40 |
| Internet/Online Media | 1440 | 3.66 |
| Kiểm toán | 1159 | 2.94 |
| Hàng không/Du lịch/Khách sạn | 371 | 0.94 |
| Thực phẩm & Đồ uống | 358 | 0.91 |
| Điện/Điện tử | 2018 | 5.12 |
| Cơ khí | 2417 | 6.14 |
| Dược Phẩm/Công nghệ sinh học | 342 | 0.87 |
| Công nghệ cao | 1001 | 2.54 |
| Bán lẻ/Bán sỉ | 2821 | 7.16 |
| Giáo dục/Đào tạo | 633 | 1.61 |
| Vận chuyển/Giao nhận | 300 | 0.76 |
| Marketing | 2413 | 6.13 |
| Bán hàng | 322 | 0.82 |
| Dịch vụ khách hàng | 3173 | 8.06 |
| Bán hàng kỹ thuật | 805 | 2.04 |
| Kế toán | 2028 | 5.15 |
| Nhân sự | 2430 | 6.17 |
| Biên phiên dịch | 960 | 2.44 |
| IT-Phần cứng/Mạng | 2153 | 5.47 |
| IT - Phần mềm | 2522 | 6.40 |
| Xuất nhập khẩu | 861 | 2.19 |
| QA/QC | 671 | 1.70 |
| Sản Xuất | 2106 | 5.35 |
| Vật Tư/Cung vận | 1012 | 2.57 |



Biểu đồ 1. Biểu đồ tỉ lệ ngành nghề việc làm thu thập được

# Download dữ liệu

<https://drive.google.com/open?id=0B6LRYpEF5kWURWtkdkk1dElNYzg>.